INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN ENGINEERING
Int. J. Numer. Meth. Engng 2001; **00**:1–6 *Prepared using* nmeauth.cls *[Version: 2000/01/19 v2.0]*

On finite element solution of the pure Neumann problem

P. Bochev[1,*,] and R. B. Lehoucq[2]

[1] *Sandia National Laboratories*[†]*, P.O. Box 5800, MS 1110, Albuquerque, NM 87185-1110.*
pbboche@sandia.gov
[2] *Sandia National Laboratories, P.O. Box 5800, MS 1110, Albuquerque, NM 87185-1110.*
rblehou@sandia.gov

SUMMARY

This paper considers the finite element approximation and algebraic solution of the pure Neumann problem. Our goal is to present a concise variational framework for the finite element solution of the Neumann problem that focuses on the interplay between the algebraic and variational problems. While many of the results that stem from our analysis are known by some experts, they are seldom derived in a rigorous fashion and remain part of numerical folklore. As a result, this knowledge is not accessible (nor appreciated) by many practitioners—both novices and experts—in one source. Our paper contributes a simple, yet insightful link between the continuous and algebraic variational forms that will prove useful. Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS: *finite elements, Neumann problem, Rayleigh-Ritz minimization, regularization, quadratic programming.*

1. INTRODUCTION

This paper is concerned with finite element solution of the pure Neumann problem

$$-\triangle u = f \text{ in } \Omega \quad \text{and} \quad \frac{\partial u}{\partial \mathbf{n}} = 0 \text{ on } \Gamma \tag{1}$$

where $\Omega \subset \mathbb{R}^N$ is a bounded open region with boundary Γ. Solutions of (1) are determined up to a constant[§] mode. The Fredholm Alternative implies that the source f must be orthogonal to this mode, that is

$$\int_\Omega f(x)dx = 0. \tag{2}$$

A direct Galerkin discretization of (1)-(2) leads to a linear system with similar properties: a stiffness matrix with a one-dimensional kernel and a source term that is orthogonal to

*Correspondence to: Sandia National Laboratories, P.O. Box 5800, MS 1110, Albuquerque, NM 87185-1110.
[†]Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.
[§]In the context of mechanical systems this mode is usually called a *rigid body motion*.

Received 1 October 2002
Revised 1 mOctober 2002

this kernel. There are two basic approaches for computing a finite element solution from this system. One, favored by some practitioners, is to constrain the candidate solution by specifying its value at a node. This eliminates the null-space and allows one to solve the linear system by a conventional (sparse) direct solver.

Alternatively, the solution can be computed from the consistent singular system either by a properly modified direct procedure that recognizes (machine) zero pivot, or a minimization based iterative solver such as conjugate gradients. This approach is less popular for three reasons: special purpose direct solvers are required, there is a general aversion towards solving singular systems, and many people are not aware that conjugate gradients work for positive semi-definite consistent linear systems.

In the extant literature, both solution techniques are formulated directly for the discrete problems without any connection to a variational problem. This situation is unsatisfactory because under closer scrutiny both approaches reveal some unsettling details. For instance, specifying solution datum at a node is inherently ambiguous, while roundoff error may render the singular system inconsistent and prevent convergence of conjugate gradients. At the same time, many well regarded FEM textbooks [2, 12, 5, 16, 8, 19, 18, 17, 11] provide only scant information on these issues. As a rule, engineering texts limit their exposition to a brief, ad hoc discussion of the first approach; see the recent textbook by Gresho and Sani [11, p.474], or the classic text [2]. Mathematically oriented finite element books, on the other hand, focus on variational problems posed in factor, or zero mean spaces [5, 10, 4], but do not discuss the practical details of implementing conforming finite element methods in these settings. As for the second approach, the solution of singular systems by the conjugate gradient algorithm is rarely discussed outside the specialized literature on iterative solvers [1, 13] or sparse direct solvers [15, 9].

The contribution of our paper is threefold. First, we seek to develop a unifying variational framework for the finite element solution of the Neumann problem that embraces existing solution techniques and presents a lucid connection between the algebraic equations and well-posed variational problems. Second, with the aforementioned connection, we present several new results that have not appeared, to best of our knowledge, in the literature. Third, we address the impact of our choices when using an iterative method of solution instead of the commonly studied impact on (sparse) direct methods for the solution of the linear system.

Since our analysis will recover widely practiced solution techniques, many of the results (and conclusions) in this paper will probably be known to an expert in mathematical theory of finite elements or an experienced practitioner of the method. Nevertheless, we feel that there is a need to provide both novices and experts with a systematic presentation of the existing body of knowledge. Moreover, our treatment reveals the common variational origins of seemingly different solution techniques, allows for their rigorous mathematical analysis and suggests new methods, a development that to the best of our knowledge has not appeared before in the extant literature.

We mention that our approach can be applied with equal success to other problems where a finite element discretization leads to a matrix with a non-trivial kernel. We have intentionally chosen to focus on the Neumann problem so as to avoid unnecessary technical detail and instead focus on the germane idea.

Finite element solution of the Neumann problem and all ensuing approaches can be completely understood by realizing that there are two variational settings that give well-posed

weak problems. Both are related to the energy functional of (1)

$$J(v,f)=\frac{1}{2}\int_\Omega |\nabla v|^2 dx-\int_\Omega fv dx \tag{3}$$

but differ in the type of optimization involved—constrained vs. unconstrained. Without constraints minimizers of (3) are determined up to a constant. The first variational setting is to factor out the constants and minimize (3) on a *factor space*. This leads to finite element methods that require solution of a singular linear system.

If we impose a suitable constraint, then (3) will have a unique minimizer in a standard Sobolev space. This is the second variational setting, and depending on how the constraint is introduced and implemented, a number of different methods follow. The standard way to enforce a constraint is to use Lagrange multipliers. We show that the popular solution method of fixing the datum at a point is simply an instance of this technique. Ultimately, solutions of finite element problems obtained by Lagrange multipliers all reduce to variations of the null-space method [14] for equality constrained quadratic programs (QPS).

A saddle-point Lagrangian formulation can also be regularized by relaxing the constraint. This leads to a class of finite element methods that have not been previously documented in the literature. Moreover, we show that these *regularized* finite element formulations have some attractive properties, especially in the context of iterative solution methods.

Throughout the paper we use the standard notation $H^s(\Omega)$ for a Sobolev space of order s with norm and inner product given by $\|\cdot\|_s$ and $(\cdot,\cdot)_s$, respectively. Seminorms on $H^s(\Omega)$ will be denoted by $|\cdot|_k$, $0\le k\le s$. For example, $|u|_1=\int_\Omega |\nabla u|^2 dx$. For $s=0$ we write $L^2(\Omega)$ instead of $H^0(\Omega)$ and denote the resulting inner product by $(\cdot,\cdot)$.

Since our study also makes use of matrix theory, we introduce some useful notation. With $\{\mathbf{e}_i\}_{i=1}^N$ and $\mathbf{I}_N$ we denote the canonical basis on $\mathbb{R}^N$ and the identity matrix of order N. For $\mathbf{x},\mathbf{y}\in\mathbb{R}^N$ the standard Euclidean norm and inner product are denoted by $\mathbf{x}^T\mathbf{y}$ and $\|\cdot\|$, respectively. The ordering of the eigenvalues of a $N\times N$ matrix $\mathbf{A}$ is $0=\lambda_1\le\lambda_2\le\cdots\le\lambda_N$.

We call attention to our specific use of bold font for matrices and vectors. Elements of matrices and vectors will be denoted by lower-case Greek letters.

2. PROJECTIONS AND INEQUALITIES

Two projection operators will play fundamental role throughout the paper. Let ω be a smooth function such that

$$(1,\omega)>0 \tag{4}$$

and consider the subspace

$$H^1_\omega(\Omega)=\{u\in H^1(\Omega)\,|\,(u,\omega)=0\} \tag{5}$$

of all functions in $H^1(\Omega)$ with zero ω-mean. For any $u\in H^1(\Omega)$ we define the operators

$$\mathcal{P}_\omega u=u-\frac{(u,\omega)}{(1,\omega)}=u-u_\omega, \tag{6}$$

where $u_\omega=(u,\omega)/(1,\omega)$ is the normalized ω-mean of u, and

$$\mathcal{P}^*_\omega u=u-\omega\frac{(1,u)}{(1,\omega)}, \tag{7}$$

respectively. A direct calculation shows that $\mathcal{P}_\omega u \in H^1_\omega(\Omega)$, $\mathcal{P}^*_\omega u \in H^1(\Omega) \cap L^2_0(\Omega)$ and that $\mathcal{P}_\omega$ and $\mathcal{P}^*_\omega$ are projectors. Therefore, $\mathcal{P}_\omega$ is a projector $H^1(\Omega) \mapsto H^1_\omega(\Omega)$ parallel to span(1) and $\mathcal{P}^*_\omega$ is a projector $H^1(\Omega) \mapsto H^1(\Omega) \cap L^2_0(\Omega)$ parallel to span(ω).

Lemma 1. *$\mathcal{P}_\omega$ and $\mathcal{P}^*_\omega$ are adjoint with respect to the L^2 inner product, that is*

$$(\mathcal{P}_\omega u, v) = (u, \mathcal{P}^*_\omega v). \tag{8}$$

Proof 1.

$$(\mathcal{P}_\omega u, v) = (u - u_\omega, v) = (u, v) - \frac{(u,\omega)}{(1,\omega)}(1, v) = (u, v - \omega(1, v)/(1, \omega)) = (u, \mathcal{P}^*_\omega v).$$

Remark 1. *We note that $\mathcal{P}^*_\omega f = 0$ for $\omega = f$ and that $\mathcal{P}_\omega$ is self-adjoint when $\omega = 1$.*

Friedrichs inequality [5, p.102] is fundamental for the analysis of the Neumann problem. A generalized version of this result follows.

Lemma 2. *Assume that Ω is simply connected and that $H^1(\Omega) \subset L^2(\Omega)$ with compact imbedding. Then, there exists a positive constant C such that*

$$\|\mathcal{P}_\omega u\|_0 \le C|u|_1 \quad \textit{and} \quad \|u\|_0 \le C(|u|_1 + |u_\omega|) \quad \textit{for every } u \in H^1(\Omega). \tag{9}$$

Proof 2. *If the first inequality in (9) is not true, then there's a sequence $\{u_k\} \subset H^1(\Omega)$ such that $\|u_k\|_1 = 1$, $u_{\omega,k} = 0$, and $|u_k|_1 < 1/k$. This sequence has a subsequence, denoted again by $\{u_k\}$ that converges weakly in $H^1(\Omega)$ and strongly in $L^2(\Omega)$ to some u. This and $|u_k|_1 < 1/k$ imply that $\nabla u = 0$ a.e. in Ω and so $u = const$ a.e. in Ω. Likewise,*

$$\int_\Omega u\omega dx = \lim_{k\mapsto\infty} \int_\Omega u_k \omega dx = 0.$$

By assumption $(1, \omega) > 0$ and so $u \equiv 0$. As a result, $u_k \mapsto 0$ in $H^1(\Omega)$, a contradiction. The second inequality follows by a similar compactness argument.

3. UNCONSTRAINED OPTIMIZATION SETTING

We consider the problem of minimizing (3) over the factor space $H^1(\Omega)/\mathbb{R}$:

$$\min_{\hat{u}\in H^1(\Omega)/\mathbb{R}} J(\hat{u}, f) \tag{10}$$

where $f \in L^2_0(\Omega)$ is given and

$$H^1(\Omega)/\mathbb{R} = \{\hat{u} \subset H^1(\Omega) \,|\, u, v \in \hat{u} \Leftrightarrow u - v = C\}. \tag{11}$$

$H^1(\Omega)/\mathbb{R}$ is a Hilbert space when equipped with the quotient norm

$$\|\hat{u}\|_{H^1(\Omega)/\mathbb{R}} = \inf_{u\in\hat{u}} \|u\|_1 \tag{12}$$

and the mapping $\hat{u} \mapsto |u|_1$, $u \in \hat{u}$ defines a norm equivalent to (12) [10, p. 13]. The Euler-Lagrange equation for (10) is to seek $\hat{u} \in H^1(\Omega)/\mathbb{R}$ such that

$$\hat{\mathcal{A}}(\hat{u}, \hat{v}) = \hat{F}(\hat{v}) \quad \forall \hat{v} \in H^1(\Omega)/\mathbb{R}, \tag{13}$$

where

$$\hat{\mathcal{A}}(\hat{u},\hat{v}) = \mathcal{A}(u,v) := \int_\Omega \nabla u \cdot \nabla v dx; \quad u \in \hat{u},\, v \in \hat{v} \tag{14}$$

is a bilinear form $H^1(\Omega)/\mathbb{R} \times H^1(\Omega)/\mathbb{R} \mapsto \mathbb{R}$, and

$$\hat{F}(\hat{u}) = F(u) := (f,u) \quad u \in \hat{u}, \tag{15}$$

is a linear functional $H^1(\Omega)/\mathbb{R} \mapsto \mathbb{R}$. Both (14) and (15) are well-defined because $\hat{\mathcal{A}}(u_1-u_2,\cdot) = \hat{\mathcal{A}}(\cdot, v_1 - v_2) = 0$ and $\hat{F}(u_1-u_2) = C\int_\Omega f dx = 0$ whenever $u_1, u_2 \in \hat{u}$, and $v_1, v_2 \in \hat{v}$. Because $|\cdot|_1$ is equivalent to the quotient norm (12), the bilinear form (14) is continuous and coercive on the quotient space. Hence (13) has a unique solution in $H^1(\Omega)/\mathbb{R}$.

4. CONSTRAINED OPTIMIZATION SETTING

To formulate a problem that has a unique minimizer out of $H^1(\Omega)$ we will require all minimizers to have a vanishing ω-mean, that is we consider the problem

$$\min_{u\in H^1(\Omega)} J(u,f) \quad \text{subject to} \quad u_\omega = 0. \tag{16}$$

The choice of ω and the handling of the constraint in (16) provide a template for all finite element methods for the Neumann problem. In view of Remark 1 $\omega = f$ will be excluded from the set of admissible weights.

4.1. A saddle-point formulation

We introduce a Lagrange multiplier $\tau \in \mathbb{R}$ and consider the saddle-point optimization problem (see problem 4.21 in [4, p 140])

$$\inf_{u\in H^1(\Omega)} \sup_{\tau\in\mathbb{R}} \left(J(u,f) + \tau u_\omega\right). \tag{17}$$

The saddle-point $(u,\tau) \in H^1(\Omega)\times\mathbb{R}$ of (17) solves the first-order optimality system

$$\begin{array}{rcccll} \mathcal{A}(u,v) & + & \tau v_\omega & = & F(v) & \forall v \in H^1(\Omega) \\ & & \sigma u_\omega & = & 0 & \forall \sigma \in \mathbb{R}. \end{array} \tag{18}$$

Theorem 1. *Problem (18) has a unique solution (u,τ) for any $f \in L^2(\Omega)$.*

Proof 3. *We apply the abstract theory of [6] and so we must show that there exists a $\gamma > 0$ for every τ so that the form $b(\tau,u) = \tau u_\omega$ satisfies the inf-sup condition*

$$\sup_{u\in H^1(\Omega)} \frac{b(\tau,u)}{\|u\|_1} \geq \gamma|\tau|.$$

We equivalently show that for a given $\tau \in \mathbb{R}$ there exists $u \in H^1(\Omega)$ such that $b(\tau,u) \geq \gamma\|u\|_1|\tau|$. Choosing $u = 1$ gives $\|u\|_1 = \sqrt{\text{meas}(\Omega)}$ and

$$b(\tau,u) = \tau(1,\omega)/(1,\omega) = \tau,$$

and so the inf-sup condition clearly holds with $\gamma = 1/\sqrt{\text{meas}(\Omega)}$. *To show that* $\mathcal{A}(\cdot,\cdot)$ *is coercive on the kernel*

$$Z = \{u \in H^1(\Omega) \,|\, b(\tau, u) = 0 \quad \forall \tau \in \mathbb{R}\}.$$

we observe that $Z = H^1_\omega(\Omega)$. *The generalized Friedrichs inequality (9) implies that* $|u|_1$ *is an equivalent norm on* $H^1_\omega(\Omega)$ *and because* $\mathcal{A}(u,u) = |u|_1^2$, *we conclude that this form is coercive on* Z. *Existence and uniqueness of a saddle-point* (u,τ) *now follows from [6].*

Restriction of (16) to Z gives the equivalent, unconstrained, *reduced* problem

$$\min_{u \in H^1_\omega(\Omega)} J(u,f). \tag{19}$$

The Euler-Lagrange equation of the reduced problem is

$$\text{seek } u \in H^1_\omega(\Omega) \text{ such that} \quad \mathcal{A}(u,v) = F(v) \quad \forall v \in H^1_\omega(\Omega). \tag{20}$$

Theorem 1 asserts that $\mathcal{A}(\cdot,\cdot)$ is coercive bilinear form on $H^1_\omega(\Omega) \times H^1_\omega(\Omega)$. Therefore, the Lax-Milgram Lemma implies that (20) is a well-posed problem for any $f \in L^2(\Omega)$.

In summary, we have the choice of either the saddle-point problem (18) or the coercive problem (20).

4.2. A stabilized saddle-point formulation

We can modify (17) by stabilizing the Lagrangian functional

$$\inf_{u \in H^1(\Omega)} \sup_{\tau \in \mathbb{R}} \left(J(u,f) + \tau u_\omega - \frac{1}{2\rho}\tau^2 \right), \tag{21}$$

where $\rho > 0$ is a stabilizing parameter. The optimality system for (21) is to seek $(u,\tau) \in H^1(\Omega) \times \mathbb{R}$ such that

$$\begin{array}{rcccll} \mathcal{A}(u,v) & + & \tau v_\omega & = & F(v) & \forall v \in H^1(\Omega) \\ & & \sigma u_\omega & = & \rho^{-1}\sigma\tau & \forall \sigma \in \mathbb{R}. \end{array} \tag{22}$$

The Lagrange multiplier can be eliminated from (22) to obtain the *regularized* problem

$$\mathcal{A}_\rho(u,v) = F(v) \quad \forall v \in H^1(\Omega), \tag{23}$$

where

$$\mathcal{A}_\rho(u,v) = \mathcal{A}(u,v) + \rho u_\omega v_\omega = \int_\Omega \nabla u \cdot \nabla v dx + \rho u_\omega v_\omega. \tag{24}$$

We remark that (23) is a first-order optimality system for the unconstrained minimization of the penalized energy functional:

$$\min_{u \in H^1(\Omega)} \left(J(u,f) + \frac{\rho}{2} u_\omega^2 \right) \equiv \min_{u \in H^1(\Omega)} J_\rho(u,f). \tag{25}$$

Theorem 2. *For every* $f \in L^2(\Omega)$ *problem (25) has a unique minimizer* $u \in H^1(\Omega)$.

Proof 4. *From (9) we see that*

$$\mathcal{A}_\rho(u,u) = |u|_1^2 + \rho u_\omega^2 \geq C\|u\|_1^2,$$

that is, (24) is coercive on $H^1(\Omega) \times H^1(\Omega)$. *Continuity of this form and* $F(\cdot)$ *are obvious and so, we can conclude that the regularized problem has a unique solution.* □

Therefore, in the present setting we can choose between the regularized saddle-point problem (22), or the coercive problem (23).

4.3. Characterization of solutions

We now consider the relationship between the solutions obtained from the constrained optimization setting and the original Neumann problem. Without stabilization we have the choice of (18) or (20), with stabilization the choice is between (21) or (23). However, within each pair the same weak solution u will be generated and so it suffices to consider the two coercive equations, that is (21) and (23).

If $f \in L^2_0(\Omega)$, both (19) and (25) have solutions that belong to a minimizing class of (10). However, (10) is not well-posed unless $f \in L^2_0(\Omega)$, while the weak problems (20) and (23) are coercive and solvable for any $f \in L^2(\Omega)$. Our next result shows that when f does not satisfy the compatibility condition (2) solutions computed by (20) and (23) solve a Neumann problem with a modified source term.

Theorem 3. *Let $\tilde{u}$ denote a solution of (23) (respectively (20)). For any $f \in L^2(\Omega)$*

$$\tilde{u}_\omega = \alpha(f, 1), \tag{26}$$

where $\alpha = 1/\rho$ for (23) and $\alpha = 0$ for (20). If $\tilde{u} \in H^2(\Omega)$, then $\tilde{u}$ solves the Neumann problem

$$-\triangle u = \mathcal{P}^*_\omega f \quad \text{in } \Omega \quad \text{and} \quad \frac{\partial u}{\partial \mathbf{n}} = 0 \quad \text{on } \Gamma.$$

Proof 5. *For $\tilde{u}$ computed by (20) formula (26) is trivially true since $\tilde{u} \in H^1_\omega(\Omega)$. To prove (26) for (23) insert $v = 1$ in (23) to obtain*

$$(f, 1) = \frac{\rho}{(1,\omega)^2}(\tilde{u}, \omega)(1, \omega) = \rho \tilde{u}_\omega.$$

Let $\tilde{u} \in H^2(\Omega)$ solve (23). Integrating (23) by parts gives

$$(-\triangle \tilde{u} - f + \omega \frac{\rho}{(1,\omega)^2}(\tilde{u}, \omega), v) + < \frac{\partial \tilde{u}}{\partial \mathbf{n}}, v >_\Gamma = 0 \quad \forall v \in H^1(\Omega).$$

From (26) $\rho(\tilde{u}, \omega)/(1, \omega)^2 = (f, 1)/(1, \omega)$, so $v \in H^1_0(\Omega)$ implies

$$-\triangle \tilde{u} - (f - \omega(f, 1)/(1, \omega)) = -\triangle \tilde{u} - \mathcal{P}^*_\omega f = 0.$$

Choosing $v \neq 0$ on Γ recovers the Neumann boundary condition.

Let $\tilde{u} \in H^2(\Omega)$ denote a solution of (20). Since $\mathcal{P}_\omega v \in H^1_\omega(\Omega)$,

$$\mathcal{A}(\tilde{u}, \mathcal{P}_\omega v) = F(\mathcal{P}_\omega v) \quad \forall v \in H^1(\Omega).$$

From the definition of $\mathcal{A}(\cdot, \cdot)$, (6), and Lemma 1

$$\mathcal{A}(\tilde{u}, \mathcal{P}_\omega v) = \mathcal{A}(\tilde{u}, v) \quad \text{and} \quad (f, \mathcal{P}_\omega v) = (\mathcal{P}^*_\omega f, v),$$

and so

$$\mathcal{A}(\tilde{u}, v) = (\mathcal{P}^*_\omega f, v) \quad \forall v \in H^1(\Omega).$$

Integrating this identity by parts gives

$$(-\triangle \tilde{u} - \mathcal{P}^*_\omega f, v) + < \frac{\partial \tilde{u}}{\partial \mathbf{n}}, v >_\Gamma = 0 \quad \forall v \in H^1(\Omega).$$

The theorem follows by first choosing $v \in H^1_0(\Omega)$ and then $v \in H^1(\Omega)$. □

Corollary 1. *If $f \in L^2_0(\Omega)$ solutions of the reduced and regularized problems coincide.*

Proof 6. *Let u^R solve (23). From (26) it is clear that $u^R_\omega = 0$ whenever $f \in L^2_0(\Omega)$, that is $u^R \in H^1_\omega(\Omega)$. Now it is easy to see that u^R also satisfies the weak problem (20).* □

5. FINITE ELEMENT SOLUTION

Throughout this section $\mathcal{T}$ denotes a uniformly regular triangulation of Ω into finite elements. For brevity we restrict attention to planar regions, triangular elements and Lagrangian finite element spaces P^k; see [4] for details. The coefficient vector of $u_h \in P^k$ with respect to a nodal basis $\{\phi_i^h\}_{i=1}^N$ is denoted by $\mathbf{u}$.

Formulation of finite element methods will be based on the link between optimization and the Neumann problem established in §§3–4. Thus, we identify finite element solution of (1) with the computation of approximate minimizers or saddle-points out of some P^k. To state the algebraic problems that will arise in the solution process we shall need the symmetric positive semi-definite stiffness matrix $\mathbf{A}$ with element i, j

$$\mathbf{A}_{i,j} = \mathcal{A}(\phi_j^h, \phi_i^h), \qquad i, j = 1, \ldots, N. \tag{27}$$

We denote the j-th column of $\mathbf{A}$ by $\mathbf{A}_j$; $\mathbf{f}_i = F(\phi_i^h)$ is the i-th element of the discrete source term $\mathbf{f}$ and $\mathbf{w}_i = (\phi_i^h, \omega)$ is the weighted basis mean vector. When $\omega = 1$ we will use $\mathbf{z}$ instead of $\mathbf{w}$. For a nodal basis $\mathbf{A}$ has a kernel spanned by the constant vector $\mathbf{c} = (1, \ldots, 1)^T$. If $\mathbf{M}$ is the mass matrix with element $\mathbf{M}_{i,j} = (\phi_j^h, \phi_i^h)$ the relationships $\mathbf{z} = \mathbf{M}\mathbf{c}$ and $(u_h, v_h) = \mathbf{u}^T\mathbf{M}\mathbf{v}$ hold. The last expression is the discrete $L^2(\Omega)$ inner product of u_h and v_h.

5.1. Finite elements in the unconstrained setting

In mathematical finite element texts the use of (13) as a well-posed weak form for the Neumann problem is standard. In contrast, this setting has found limited acceptance among practitioners because formally it requires a finite element subspace $P^k/\mathbb{R}$ of $H^1(\Omega)/\mathbb{R}$, formulation of the ensuing method is never clarified, and the matrix problem is singular. However, the ambiguities of a factor space setting can be easily avoided within the optimization framework. Since $P^k/\mathbb{R}$ is isomorphic to $\mathbb{R}^N/(\ker(\mathbf{A}) \equiv \mathbf{c})$ the discrete equivalent of (10) and its algebraic form are

$$\min_{\hat{u}^h \in P^k/\mathbb{R}} J(\hat{u}^h, f) \equiv \min_{\hat{\mathbf{u}} \in \mathbb{R}^N/\mathbf{c}} \frac{1}{2}\hat{\mathbf{u}}^T\mathbf{A}\hat{\mathbf{u}} - \hat{\mathbf{u}}^T\mathbf{f} \tag{28}$$

Therefore, a finite element method in the factor space setting simply amounts to computation of an arbitrary member from the minimizing class $\hat{\mathbf{u}}^h$. Such a member can be determined by solving the linear system

$$\mathbf{A}\hat{\mathbf{u}} = \mathbf{f} \tag{29}$$

by a sparse direct method modified so that a zero pivot can be detected. However, floating point arithmetic complicates this decision because the solver needs to decide when a pivot is negligible. Instead we recommend an iterative scheme be applied directly to (28). Indeed, as long as the $\mathbf{f}$ is in the range of $\mathbf{A}$ the quadratic functional in (28) has a finite lower bound. As a result, the conjugate gradient algorithm will generate a minimizing sequence that converges modulo $\ker(\mathbf{A})$; see Theorem 13.11, [1, p. 583]. The rate of convergence of the conjugate gradient algorithm depends on the ratio $\kappa_\epsilon(\mathbf{A}) = \lambda_N(\mathbf{A})/\lambda_2(\mathbf{A})$ or the effective condition number.

An important practical consideration for (29) is that the discrete source $\mathbf{f}$ must be discretely orthogonal to the constant vector $\mathbf{c}$ and $\mathbf{A}\mathbf{c} = \mathbf{0}$. Since $(1, f) = \mathbf{c}^T\mathbf{f}$ in exact arithmetic, the linear system will be consistent whenever the Neumann problem is solvable, that is when f has zero mean. In practice the source $\mathbf{f}$ and the matrix $\mathbf{A}$ are computed in floating point

Table I. Loss of consistency in (29) within CG. $\mathbf{x}^{(j)}$ denotes the CG solution at the j-th iteration.

P^1 elements - nonuniform				P^2 elements - nonuniform		
Quad.	$\mathbf{c}^T\mathbf{f}$	iter	$\|\mathbf{f}-\mathbf{A}\mathbf{x}^{(j)}\|$	$\mathbf{c}^T\mathbf{f}$	iter	$\|\mathbf{f}-\mathbf{A}\mathbf{x}^{(j)}\|$
1	-2.522E-03	1000	0.38000	-1.857E-03	underintegrated $\mathbf{A}$	
3	3.069E-05	208	0.1242E-05	1.610E-04	1000	0.2128E+02
7	-2.744E-09	85	0.1329E-05	-2.023E-08	169	0.8327E-06
P_1 elements - uniform				P_2 elements - uniform		
3	4.628E-15	55	0.9786E-05	-3.123E-16	54	0.7439E-06

arithmetic via quadrature. As a result, $\mathbf{c}^T\mathbf{f}$ equals $(1, f)$ only approximately and (29) may become inconsistent. To restore consistency we take a cue from Theorem 3 and introduce the discrete projector $(\mathbf{P}^T\mathbf{f})_i \equiv (\mathcal{P}^*_\omega f, \phi^h_i)$, $i = 1, \ldots, N$. A direct calculation shows that

$$\mathbf{P}^T = \mathbf{I} - \frac{\mathbf{w}\mathbf{c}^T}{\mathbf{w}^T\mathbf{c}}.$$

Application of the projector to the linear system results in

$$(\mathbf{P}^T\mathbf{A}\mathbf{P})\mathbf{u} = \mathbf{P}^T\mathbf{f}. \tag{30}$$

The matrix $\mathbf{P}$ is the discrete analogue of the projector $\mathcal{P}_\omega$ and so the FEM solution $\mathbf{P}\mathbf{u}$ has zero ω-mean, that is $\mathbf{w}^T\mathbf{P}\mathbf{u} = 0$. We remark that the iterative solution of semi-definite systems and application of projectors is rarely discussed beyond specialized texts on iterative solvers and does not seem to be widely known among finite element practitioners. This is another reason for the limited use of (29).

Let us demonstrate that the use of a projector to maintain consistency of (29) is not unfounded, especially for unstructured meshes. To test effects of numerical quadrature we consider the zero mean source f defined by evaluating (1) at $u(x, y) = \cos(\pi x^2)\cos(2\pi y)$ on the unit square. We solve (29) with discrete sources $\mathbf{f}$ computed using linear (1 point), quadratic (3 point) and quintic (7 point) quadrature rules [7, p.343].

Table I shows that for P^2 elements on non-uniform meshes, the 3 point rule leads to a numerically inconsistent linear system and so the conjugate gradient algorithm diverges. For nonuniform P^1 elements the 3 point rule does suffice but requires 2.5 times more conjugate gradient iterations than the 7 point rule.

On uniform grids all three quadrature rules led to a discrete source $\mathbf{f}$ with *exact* zero mean and a consistent (to machine precision) linear system. Table I shows that in this case conjugate gradients converged without a difficulty. This contrasting behavior clearly demonstrates the importance of maintaining consistency in (29).

5.2. Finite elements in the constrained setting

The starting point now is the constrained problem (16). To define a finite element solution we restrict minimization of (16) to a subspace P^k of $H^1(\Omega)$ and note that $u_{h,\omega} = 0$ if and only if

$\mathbf{u}^T\mathbf{w} = 0$. As a result, the discrete equivalent of (16) and its algebraic form are

$$\min_{\substack{u_h \in P^k \\ u_{h,\omega}=0}} J(u_h, f) \equiv \min_{\substack{\mathbf{u}\in\mathbb{R}^N \\ \mathbf{w}^T\mathbf{u}=0}} \frac{1}{2}\mathbf{u}^T\mathbf{A}\mathbf{u} - \mathbf{u}^T\mathbf{f}. \tag{31}$$

In the optimization literature (31) is known as an equality constrained quadratic program [14]. This quadratic program can be solved in a number of ways. In all cases however, we are led to an algebraic equation that is related to one of the four variational problems (18), (20), (21), or (23). In what follows we consider the variational settings of §4.1 and §4.2 and demonstrate their relationship with (31).

5.2.1. The saddle-point formulation The algebraic equivalent of saddle-point equation (18) is a symmetric, indefinite linear system.

$$\begin{pmatrix} \mathbf{A} & (\mathbf{w}^T\mathbf{c})^{-1}\mathbf{w} \\ (\mathbf{w}^T\mathbf{c})^{-1}\mathbf{w}^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \tau \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ 0 \end{pmatrix} \tag{32}$$

This system can be obtained directly from (31) by introducing a Lagrange multiplier for the algebraic constraint. The matrix in (32) is called the Karush-Kuhn-Tucker (KKT) matrix. One way to compute a finite element approximation is to solve (32) by either a sparse direct method or an iterative method. Another approach that exploits the structure in the KKT matrix is the *null-space* method. The alternative *range-space* method requires that $\mathbf{A}$ is nonsingular and is not applicable to (32).

The constraint $\mathbf{w}^T\mathbf{u} = 0$ implies that the minimizer belongs to the space $\text{span}(\mathbf{w})^\perp$ or, equivalently, the null-space of $\mathbf{w}^T$. Let $\mathbf{B} \in \mathbb{R}^{N\times(N-1)}$ denote a matrix whose columns form a basis for $\text{span}(\mathbf{w})^\perp$. Then $\mathbf{u} = \mathbf{B}\mathbf{v}$ and (31) is equivalent to an unconstrained problem

$$\min_{\mathbf{v}\in\mathbb{R}^{N-1}} \frac{1}{2}\mathbf{v}^T\mathbf{B}^T\mathbf{A}\mathbf{B}\mathbf{v} - \mathbf{v}^T\mathbf{B}^T\mathbf{f} \tag{33}$$

in terms of $\mathbf{v}$. The null-space method for (32) amounts to constructing the matrix $\mathbf{B}$ and solving the symmetric positive definite linear system

$$\mathbf{B}^T\mathbf{A}\mathbf{B}\mathbf{v} = \mathbf{B}^T\mathbf{f}. \tag{34}$$

The null-space method is the variational equivalent of "minimization on the kernel" that gave the reduced problem (19). Let us show that conforming discretization of (19) is in turn equivalent to an explicit method for constructing the matrix $\mathbf{B}$. For this purpose we restrict (19) to a finite element subspace $P^k_\omega = P^k \cap H^1_\omega(\Omega)$ of $H^1_\omega(\Omega)$. Since P^k_ω is isomorphic with $\mathbb{R}^{N-1}$, the discrete minimization problem and its algebraic form are

$$\min_{u_h\in P^k_\omega} J(u_h, f) \equiv \min_{\mathbf{v}\in\mathbb{R}^{N-1}} \frac{1}{2}\mathbf{v}^T\mathbf{A}_\omega\mathbf{v} - \mathbf{v}^T\mathbf{f}_\omega, \tag{35}$$

where $\mathbf{A}_\omega$, $\mathbf{f}_\omega$, and $\mathbf{v}$ denote a stiffness matrix, right hand side and a coefficient vector relative to some basis $\{\psi_i\}_{i=1}^{N-1}$ of P^k_ω. Let $\mathbf{B}$ denote the transformation matrix between this basis and the standard nodal basis of P^k. If $\mathbf{u}$ contains the nodal coefficients of u_h relative to P^k and $\mathbf{v}$ are the coefficients of this function relative to the basis in P^k_ω then

$$\mathbf{u} = \mathbf{B}\mathbf{v} \quad \text{and} \quad \mathbf{A}_\omega = \mathbf{B}^T\mathbf{A}\mathbf{B}.$$

Because $\mathcal{A}(\cdot,\cdot)$ is coercive on $H^1_\omega(\Omega)\times H^1_\omega(\Omega)$ and $P^k_\omega \subset H^1_\omega(\Omega)$ the matrix $\mathbf{A}_\omega$ is symmetric and positive definite.

In general, $\{\psi_i\}_{i=1}^{N-1}$ need not be a nodal basis. In this case the coefficients in $\mathbf{v}$ are linear combinations of the nodal values in $\mathbf{u}$. Because nodal bases are easier to work with let us show how one can be constructed for a given weight ω. Suppose that¶ $(\phi^h_\ell,\omega)\neq 0$ for some ℓ between 1 and N. Solving $(u_h,\omega)\equiv\sum_{i=1}^N \alpha_i(\phi^h_i,\omega)=0$, for the ℓth term gives the set of functions

$$\psi^h_{i,\ell}=\phi^h_i-\phi^h_\ell\frac{(\phi^h_i,\omega)}{(\phi^h_\ell,\omega)}\quad i=1,\ldots,N;\quad i\neq\ell \tag{36}$$

parameterized by ℓ, and a space $P^k_\omega=\mathrm{span}\{\psi^h_{i,\ell}\}_{i\neq\ell}\subset H^1_\omega(\Omega)$. Note that P^k_ω does not have a degree of freedom associated with the (arbitrarily chosen) triangulation node x_ℓ. By definition $(\psi^h_{i,\ell},\omega)=0$, and $\psi^h_{i,\ell}(x_j)=\delta_{ij}$ and so (36) is a nodal basis. The following result explains how the transformation matrix $\mathbf{B}$ can be constructed without forming explicitly the basis (36).

Theorem 4. *The transformation matrix for the basis (36) is*

$$\mathbf{B}_{\ell,\omega}=\Big(\mathbf{I}-\frac{\mathbf{e}_\ell\mathbf{w}^T}{\mathbf{e}_\ell^T\mathbf{w}}\Big)\mathbf{I}^\ell_N$$

where $\mathbf{I}^\ell_N$ denotes a unit matrix with deleted ℓ^{th} row and column.

Proof 7. *Given a node $x_\ell\in\mathcal{T}_h$ the entries of $\mathbf{A}_\omega$ are*

$$\begin{aligned}\mathbf{A}_\omega(i,j)\quad&\equiv\quad \mathcal{A}(\psi^h_{j,\ell},\psi^h_{i,\ell})=\mathcal{A}(\phi^h_j,\phi^h_i)-\frac{(\phi^h_i,\omega)}{(\phi_\ell,\omega)}\mathcal{A}(\phi^h_j,\phi^h_\ell)-\frac{(\phi^h_i,\omega)}{(\phi_\ell,\omega)}\mathcal{A}(\phi^h_\ell,\phi^h_i)\\&+\quad\frac{(\phi^h_i,\omega)(\phi^h_j,\omega)}{(\phi_\ell,\omega)^2}\mathcal{A}(\phi^h_\ell,\phi^h_\ell).\end{aligned}\tag{37}$$

Since $\mathcal{A}(\phi^h_j,\phi^h_i)=\mathbf{A}_{i,j}$ and $(\phi^h_i,\omega)=\mathbf{w}_i$,

$$\mathbf{A}_\omega=(\mathbf{I}^\ell_N)^T\Big(\mathbf{A}-\frac{1}{\mathbf{e}_\ell^T\mathbf{w}}(\mathbf{w}\mathbf{A}_j^T+\mathbf{A}_j\mathbf{w}^T)+\frac{\mathbf{A}_{i,j}}{(\mathbf{e}_\ell^T\mathbf{w})^2}\mathbf{w}\mathbf{w}^T\Big)\mathbf{I}^\ell_N=\mathbf{B}^T_{\ell,\omega}\mathbf{A}\mathbf{B}_{\ell,\omega}\quad\square. \tag{38}$$

Consider now a situation where $\mathbf{w}=\mathbf{e}_\ell$ so that the constraint in (31) is $\mathbf{e}_\ell^T\mathbf{u}=0$. In this case Theorem 4 gives the transformation matrix as

$$\mathbf{B}=\Big(\mathbf{I}-\mathbf{e}_\ell\mathbf{e}_\ell^T\Big)\mathbf{I}^\ell_N=\mathbf{I}^\ell_N.$$

Therefore, $\mathbf{A}_\omega$ is simply $\mathbf{A}$ with deleted ℓ^{th} row and column. Note that $\mathbf{e}_\ell^T\mathbf{u}=0$ is the same as $u_h(x_\ell)=0$ and so this is simply the standard method of specifying the solution value at a node and is a variant of the null-space method.

Our framework allows us to establish an interesting link between the linear system and a variational equation. Let ϕ^h_ℓ denote the basis function associated with node x_ℓ in some triangulation $\mathcal{T}_h$, and consider a weight function $\omega_{h,\ell}$ such that

$$(\phi^h_\ell,\omega_{h,\ell})=1\quad\text{and}\quad(\phi^h_k,\omega_{h,\ell})=0\quad\text{for } k\neq\ell. \tag{39}$$

¶This assumption is necessary because Lagrangian basis functions may have zero mean. One example is given by the P^2 basis functions associated with the nodes of a triangulation.

Then $\mathbf{w} = \mathbf{e}_\ell$ and fixing the solution value can be viewed as a conforming discretization of the saddle-point (18) or the reduced (20) problems with a constraint given by

$$(\omega_{h,\ell}, u) = 0.$$

While the choice of $\omega_{h,\ell}$ is not unique, (39) formally implies that $\omega_{h,\ell} \mapsto \delta(x_\ell)$ as $\mathcal{T}_h$ is refined. Because the delta function is in the dual of $H^1(\Omega)$ only in one dimension, this constraint will become ill-posed in two and three dimensions as $h \to 0$. We conclude that specifying the solution at a node leads to a ill-posed variational problem in two and three dimensions.

This, and the arbitrary choice of x_ℓ raises two questions about the resulting algebraic systems that, to the best of our knowledge, have not been yet discussed in the literature. The first, and more obvious question, is whether or not the choice of x_ℓ has any impact on the conditioning of the linear systems. The second, more subtle question, is how these systems behave asymptotically as $h \to 0$. The answer to the first question is important because a customary choice for x_ℓ is usually at the corners of Ω and so we want to know whether or not this is the best possible location. The relevance of the second question stems from the degeneration of the associated variational problem as $h \to 0$. To address these issues we recall a standard finite element result.

Lemma 3. *Let V^h denote an H^1 conforming finite element space and assume that there exists a constant C_P^h such that a discrete Poincaré inequality ´*

$$C_P^h \|u_h\|_0 \leq \|\nabla u_h\|_0 \tag{40}$$

holds for any $u_h \in V^h$. Then, there exists $C > 0$ such that

$$cond(\mathbf{A}_h) \leq h^{-2} \frac{C}{C_P^h}, \tag{41}$$

where $\mathbf{A}_h$ is generated by $\mathcal{A}(\cdot,\cdot)$.

The discrete Poincaré inequality that is relevant to our discussion is established in the next Theorem. For brevity we consider the case when Ω is the unit cube in $\mathbb{R}^d$ where $d = 1, 2, 3$.

Theorem 5. *Let $\Omega = (0,1)^d$ and $\Gamma_D \subseteq \partial\Omega$ be the Dirichlet portion of the boundary. If $meas(\Gamma_D) > 0$ and $u_h = 0$ on Γ_D then,*

$$C_P^h \|u_h\|_0 \leq |u_h|_1 \tag{42}$$

where

$$C_P^h = \left(\frac{\mu}{2^{d-1}}\right)^{1/2} \quad and \quad \mu = \begin{cases} meas(\Gamma_D) & for\ d > 1 \\ 1 & for\ d = 1 \end{cases}$$

Proof 8. *Without loss of generality, $\Gamma_D = \{(0,y)\,|\,0 \leq y \leq y_D\}$ and let $\mathbf{x} = (x,y) \in \Omega$ be arbitrary. For $\hat{y} \leq y_D$ the points $\mathbf{x}$, $\mathbf{x}_1 = (x,\hat{y})$ and $\mathbf{x}_D = (0,\hat{y})$ form a path from $\mathbf{x}$ to Γ_D. Because $u_h \in H^1(\Omega) \cap C^0(\Omega)$, and $u_h(\mathbf{x}_D) = 0$*

$$u_h(x,y) = u_h(x,\hat{y}) + \int_{\hat{y}}^{y} u_{h,y}(x,\eta) d\eta.$$

Squaring both sides, using Cauchy's inequality and integrating along Γ_D gives

$$meas(\Gamma_D) {u_h}^2(x,y) \leq 2 \int_{\Gamma_D} u_h^2(x,\eta) d\eta + 2\, meas(\Gamma_D) \int_0^1 u_{h,y}^2(x,\eta) d\eta.$$

To estimate the right hand side note that

$$u(x,\hat{y}) = u_h(\mathbf{x}_1) = u(\mathbf{x}_D) + \int_0^x u_{h,x}(\xi,\hat{y})d\xi = \int_0^x u_{h,x}(\xi,\hat{y})d\xi.$$

Cauchy's inequality and integration along Γ_D *give*

$$\int_{\Gamma_D} u_h^2(x,\eta)d\eta \leq \int_\Omega u_{h,x}^2(\xi,\eta)d\xi d\eta$$

and therefore,

$$meas(\Gamma_D)u_h^2(x,y) \leq 2\left[\int_\Omega u_{h,x}^2(x,y)dxdy + meas(\Gamma_D)\int_0^d u_{h,y}^2(x,\eta)d\eta\right].$$

Integrating both sides over Ω *and noting that* $meas(\Omega) = 1$ *completes the proof in two dimensions. The other cases follow in a similar manner.*

Theorem 5 proves fundamental for the understanding of the asymptotic behavior of the matrices obtained by deleting a row and a column in the singular matrix $\mathbf{A}$. Suppose that $d > 1$ and $\mathbf{x}_\ell \in \partial\Omega$. Therefore, the linear system (34) is formally associated with a mixed boundary value problem where $\Gamma_D = \text{supp}(\omega_{h,\ell}) \cap \partial\Omega$. Because $\omega_{h,\ell}$ approaches a delta function located at $\mathbf{x}_\ell$ then $\text{meas}(\Gamma_D) = O(h^{d-1})$ easily follows. As a result,

$$\text{cond}(\mathbf{A}_{\omega_{h,\ell}}) \leq h^{-2}\frac{C}{C_P^h} = O(h^{-\frac{3+d}{2}}). \tag{43}$$

Corollary 2. *Let* $\mathbf{x}_M$ *denote the center of* Ω. *If*

$$\Gamma_D = \{\mathbf{x} = (\xi^1,\ldots,\xi^d) \in \Omega \,|\, \xi^i \in [\xi_M^i - \alpha, \xi_M^i + \alpha], i \neq k\},$$

that is, Γ_D *lies on the center of a plane perpendicular to the* k*th coordinate direction that coincides with* $\mathbf{x}_M$, *then (42) holds with*

$$C_M^h = 2C_P^h.$$

Proof 9. *We apply (42) to the* 2^N *identical subdomains* Ω_k *obtained by cutting* Ω *through* $\mathbf{x}_M$ *by planes perpendicular to the* N *coordinate directions. The side length of each* Ω_k *equals* $d/2$ *and* $meas(\Gamma_D \cap \partial\Omega_k) = meas(\Gamma_D)/2^{N-1}$ *As a result, (42) holds on each* Ω_k *with*

$$\left(C_M^h\right)^2 = \frac{2^{1-d}\mu}{2^{d-1}(1/2)^{d+1}} = \frac{2^{d+1}}{2^{d-1}}\frac{\mu_M}{2^{d-1}} = \left(2C_P^h\right)^2.$$

The proof follows by summing up (42) for all subdomains.

This corollary reveals that the optimal location for $\mathbf{x}_\ell$ is the barycenter of Ω. The ratio between the Poincaré constants when $\Gamma_D \subset \partial\Omega$ and when $\mathbf{x}_M \in \Gamma_D$ is 2. As a result, fixing the solution value at the boundary can lead to a condition number that is up to four times larger than when the solution is fixed at $\mathbf{x}_M$. In one dimension, on a uniform grid, this bound is sharp and can be established by Fourier analysis.

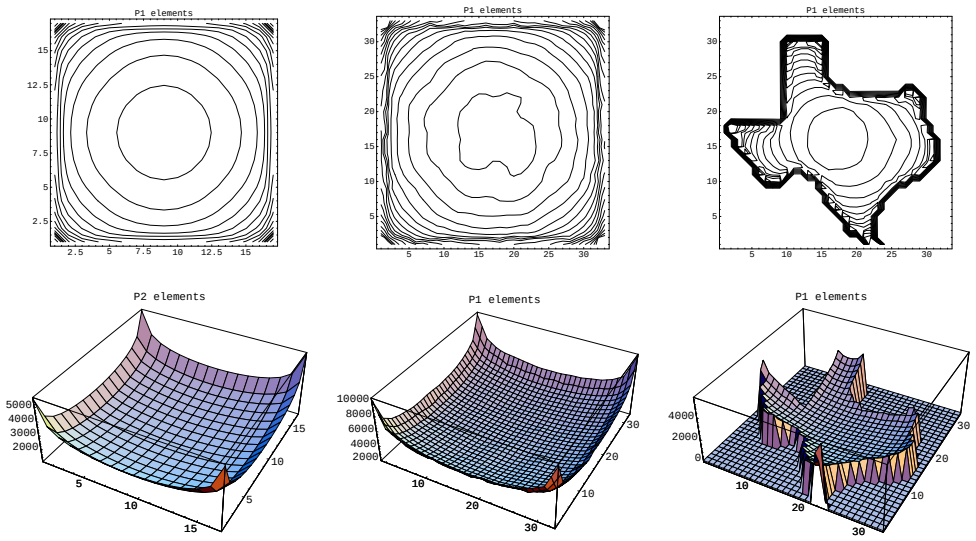

Figure 1. $\kappa(\mathbf{A}_{\omega_{h,\ell}})$ for P^1 elements: uniform and nonuniform grids on $[0,1]^2$ and nonuniform grid for Texas shaped region.

Lemma 4. *Let $N = 1$, and $\mathbf{A}_{\omega_{h,D}}$, $\mathbf{A}_{\omega_{h,M}}$ denote the matrices corresponding to solution value fixed at an endpoint and the barycenter $\mathbf{x}_M$, respectively. Then*

$$\frac{cond(\mathbf{A}_{\omega_{h,D}})}{cond(\mathbf{A}_{\omega_{h,M}})} = 4 + O(h)$$

For the proof of this lemma we refer to [3]. In more than one dimension optimality of the barycenter can be confirmed numerically. Figure 1 shows plots of the condition number of $\mathbf{A}_{\omega_{h,\ell}}$ as a function of the node location for three different triangulations and P^1 elements. In all three cases the condition number is minimized when $\mathbf{x}_\ell$ is near or at the centroid of the region, while the corners lead to the highest condition number. These patterns were observed [3] for other regions and elements, including P^2 elements.

5.2.2. The stabilized saddle-point formulation The linear system

$$\begin{pmatrix} \mathbf{A} & (\mathbf{w}^T\mathbf{c})^{-1}\mathbf{w} \\ (\mathbf{w}^T\mathbf{c})^{-1}\mathbf{w}^T & (-\rho)^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \tau \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ 0 \end{pmatrix} \tag{44}$$

is the algebraic equivalent of the stabilized saddle-point problem (21). As in §4.2, the Lagrange multiplier can be eliminated to obtain a system only in terms of $\mathbf{u}$:

$$\mathbf{A}_\rho \mathbf{u} \equiv \left(\mathbf{A} + \frac{\rho}{(\mathbf{w}^T\mathbf{c})^2} \mathbf{w}\mathbf{w}^T \right) \mathbf{u} = \mathbf{f}. \tag{45}$$

This equation is the necessary condition for the quadratic program

$$\min_{u_h \in P^k} J_\rho(u_h, f) \equiv \min_{\mathbf{u} \in \mathbb{R}^N} \frac{1}{2}\mathbf{u}^T\mathbf{A}\mathbf{u} - \mathbf{u}^T\mathbf{f} + \rho\frac{(\mathbf{w}^T\mathbf{u})^2}{(\mathbf{w}^T\mathbf{c})^2}, \tag{46}$$

Table II. CG solution of (45). $\mathbf{x}^{(j)}$ denotes the CG solution at the j-th iteration.

P_1 elements - nonuniform				P_2 elements - nonuniform		
Quad.	$\mathbf{c}^T\mathbf{f}$	iter	$\|\mathbf{f}-\mathbf{A}\mathbf{x}^{(j)}\|$	$\mathbf{c}^T\mathbf{f}$	iter	$\|\mathbf{f}-\mathbf{A}\mathbf{x}^{(j)}\|$
1	-2.522E-03	85	0.1358E-05	-1.857E-03	underintegrated $\mathbf{A}$	
3	3.069E-05	85	0.1242E-05	1.610E-04	169	0.8667E-06
7	-2.744E-09	85	0.1329E-05	-2.023E-08	169	0.8327E-06

that is a discrete counterpart of (25). From Theorem 2 it follows that $\mathbf{A}_\rho$ is symmetric and positive definite. The sparsity pattern of $\mathbf{A}_\rho$ depends on the choice of ω because $\mathbf{A}_\rho$ is a rank-one correction of the singular matrix $\mathbf{A}$. If support of ω overlaps with only a few elements in $\mathcal{T}_h$, the vector $\mathbf{w}$ will have only a few non-zero entries and $\mathbf{A}_\rho$ will have a sparsity pattern similar to that of $\mathbf{A}$. In this case a sparse direct solver can be used.

When the support of ω is larger, for instance if $\omega = 1$, then $\mathbf{w}\mathbf{w}^T$ is dense and formally, $\mathbf{A}_\rho$ is also dense. While a direct elimination is not practical in this case, (45) can be solved iteratively for almost the same cost as (29). Typically, an iterative solver requires one matrix vector product $\mathbf{A}_\rho\mathbf{u}$ per iteration. This product can be computed by

1. forming the vector $\mathbf{v} = \mathbf{A}\mathbf{u}$;
2. computing the scalar $\mu = \rho(\mathbf{w}^T\mathbf{u})$;
3. updating $\mathbf{v} + \mu\mathbf{w}$.

Step 1 is standard part of any finite element solver, so the only additional work involved is the dot product in step 2 ($2N-1$ flops) and the update in step 3 ($2N$ flops). The row vector $\mathbf{w}^T$ can be precomputed and stored rendering the computation of μ efficient.

Theorem 2 also implies that the regularized system (45) must be solvable for any discrete source $\mathbf{f}$. This means that iterative solver performance should not degrade as in Table I for low order quadrature. Table II contains convergence history for Jacobi preconditioned conjugate gradients applied to (45) and the same exact solution as in Section 5.1. Regardless of the quadrature we see identical convergence of the solver.

The following theorem proves fundamental for understanding the structure of $\mathbf{A}_\rho$ and how the rank-one update modifies the null-space of $\mathbf{A}$.

Theorem 6. *Let $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$ denote the eigendecomposition of the singular stiffness matrix $\mathbf{A}$, with $\mathbf{Q}\mathbf{e}_1 = N^{-1/2}\mathbf{c}$ and $\mathbf{A}\mathbf{c} = \mathbf{0}$. If $\mathbf{w} = \|\mathbf{w}\|\cos(\phi)\mathbf{c} + \mathbf{r}$, where $\mathbf{r}^T\mathbf{c} = 0$ and ϕ measures the positive angle between $\mathbf{c}$ and $\mathbf{w}$, then*

$$\|\mathbf{A}_\rho - \mathbf{Q}(\mathbf{\Lambda} + \rho\|\mathbf{w}\|^2\cos^2(\phi)\mathbf{e}_1\mathbf{e}_1^T)\mathbf{Q}^T\| \le \rho\|\mathbf{w}\|^2(\sin(2\phi) + \sin^2(\phi)). \tag{47}$$

Proof 10. *From the identity*

$$\mathbf{A}_\rho = \mathbf{A} + \rho\mathbf{w}\mathbf{w}^T = \mathbf{Q}(\mathbf{\Lambda} + \rho(\mathbf{Q}^T\mathbf{w})(\mathbf{Q}^T\mathbf{w})^T)\mathbf{Q}^T$$

and the hypothesis on $\mathbf{Q}\mathbf{e}_1$, we have

$$\mathbf{Q}^T\mathbf{w} = \mathbf{Q}^T(\|\mathbf{w}\|\cos(\phi)\mathbf{c} + \mathbf{r}) = \|\mathbf{w}\|\cos(\phi)\mathbf{e}_1 + \mathbf{Q}^T\mathbf{r}$$

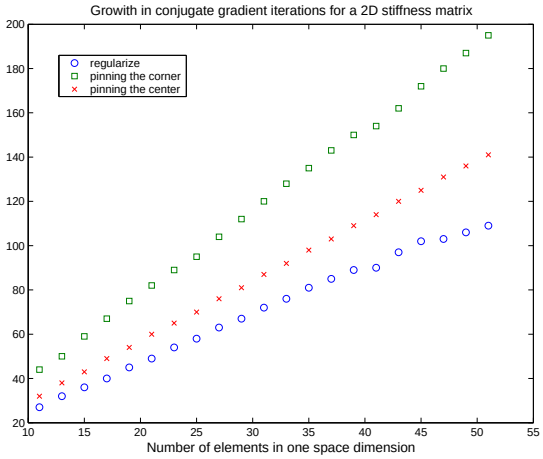

Figure 2. Growth in conjugate gradient iterations for the solution of (45) vs. (34) for a stiffness matrix from a bilinear quadrilateral approximation of a two dimensional problem. Two choices of $\mathbf{B} = \mathbf{I}_N^\ell$ corresponding to specifying the center and corner nodal coefficient are used.

and hence

$$\begin{aligned}\mathbf{Q}^T\mathbf{w}(\mathbf{Q}^T\mathbf{w})^T &= \Big(\|\mathbf{w}\|\cos(\phi)\mathbf{e}_1 + \mathbf{Q}^T\mathbf{r}\Big)\Big(\|\mathbf{w}\|\cos(\phi)\mathbf{e}_1 + \mathbf{Q}^T\mathbf{r}\Big)^T \\ &= \|\mathbf{w}\|^2\cos^2\phi\mathbf{e}_1\mathbf{e}_1^T + \|\mathbf{w}\|\cos(\phi)\mathbf{Q}^T\mathbf{r}\mathbf{e}_1^T + \|\mathbf{w}\|\cos(\phi)\mathbf{e}_1\mathbf{r}^T\mathbf{Q} + \mathbf{Q}^T\mathbf{r}(\mathbf{Q}^T\mathbf{r})^T.\end{aligned}$$

The theorem now easily follows by noting that $\|\mathbf{Q}^T\mathbf{r}(\mathbf{Q}^T\mathbf{r})^T\| = \|\mathbf{r}\|^2$, $\|\mathbf{Q}^T\mathbf{r}\mathbf{e}_1^T\| = \|\mathbf{r}\|\|\mathbf{e}_1\| = \|\mathbf{r}\|$, *and* $\|\mathbf{r}\| = \|\mathbf{w}\|\sin(\phi)$. □

This theorem shows that with a proper choice of ρ the rank-one update modifies the zero eigenvalue of $\mathbf{A}$ to a positive one and only perturbs the eigenvectors. Furthermore,

$$\mathbf{A}_\rho\mathbf{c} = \rho\|\mathbf{w}\|^2\cos^2(\phi)\mathbf{c} + \rho\|\mathbf{w}\|\cos(\phi)\mathbf{r},$$

that is, the constant mode of $\mathbf{A}$ is an approximate eigenvector of $\mathbf{A}_\rho$. Moreover, if ρ is between λ_2 and λ_N then condition number of $\mathbf{A}_\rho$ equals the effective condition number of $\mathbf{A}$.

Recall that (43) implies condition numbers higher than the effective condition number whenever solution is being specified at a point. This can be confirmed by comparing the conjugate gradient convergence of (45) and (34) when $\mathbf{B} = \mathbf{I}_N^\ell$. Figures 2-3 show the results when the pure Neumann problem is solved on the unit square and on the unit cube by bilinear and trilinear finite elements, respectively. The zero mean sources $\triangle u(x,y) = \triangle\cos(\pi x^2)\cos(2\pi y)$ and $\triangle u(x,y,z) = \triangle\cos(\pi x^2)\cos(2\pi y)\cos(z^3\pi)$ are used. The choices of $\mathbf{B} = \mathbf{I}_N^\ell$ correspond to specifying the center and the corner nodal coefficients. Figures 2-3 reveal a substantial and growing gap between the iteration counts for the regularized approach and that of specifying the solution at a point. Figure 4, on the other hand, shows that with respect to the mesh size this gap grows faster in three dimensions, and hence supports the conclusion of (43) and Theorem 5.

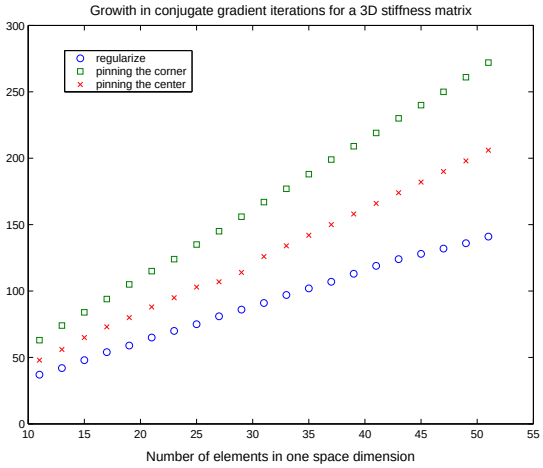

Figure 3. Growth in conjugate gradient iterations for the solution of (45) vs. (34) for a stiffness matrix from a trilinear quadrilateral approximation of a three dimensional problem. Two choices of $\mathbf{B} = \mathbf{I}_N^\ell$ corresponding to specifying the center and corner nodal coefficient are used.

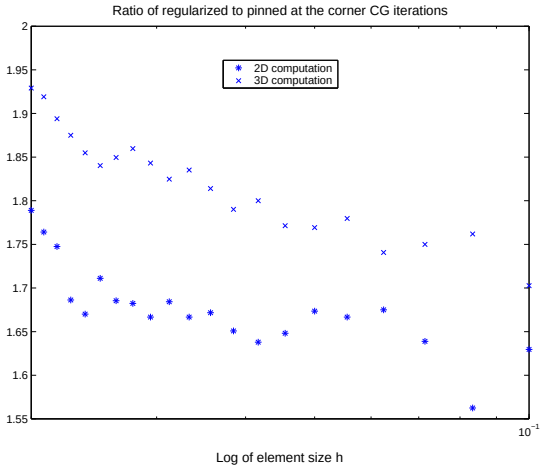

Figure 4. Ratio of the number of regularized to pinned at the center conjugate gradient iterations.

6. CONCLUSIONS

We demonstrated that finite element methods for the Neumann problem originate from two optimization settings. The first requires minimization of a quadratic energy functional on a factor space and leads to singular linear systems. These systems can be solved iteratively provided consistency is maintained by a discrete projector to ensure that the source remains discretely orthogonal to the constant mode.

The second optimization setting involves constrained minimization of a quadratic functional

and leads to an equality constrained quadratic program. The manner in which the constraint is treated defines yet another two classes of finite element methods, while the choice of the constraint describes the different methods within each class.

The first class corresponds to the application of the null-space method for the solution of the quadratic program. The method of specifying a solution value at a node is an instance of this class. Moreover, we established that this method can be associated with a variational formulation involving a weight function approaching a delta function as $h \to 0$. As a result, condition numbers of the resulting matrices are larger than the effective condition number of the singular matrix. Our analysis also indicates that the optimal location for the fixed value node is at the barycenter of Ω, a result that is also confirmed by numerical experiments.

The second class of finite element methods corresponds to a regularized formulation of the constrained minimization problem. Here we were led to a new class of methods for the Neumann problem that provide symmetric positive definite linear systems with effective condition numbers. Moreover, the sparsity pattern of the rank one update can be controlled so as to match the sparsity pattern of the singular matrix by taking a weight function with the appropriate support. We recommend the regularized method whenever an iterative solution method is used to compute the finite element solution.

ACKNOWLEDGEMENT

We would like to thank Doug Arnold, Martin Berggren, Quang Du and Michael Pernice for helpful discussions and Ulrich Hetmaniuk for generating matlab code for the matrices and source terms used in the experiments of the last section.

REFERENCES

1. O. Axelsson. *Iterative Solution methods*. Cambridge University Press, Cambridge, 1994.
2. E. Becker, G. Carey, and T. Oden. *Finite Elements, an Introduction*, volume 1. Prentice-Hall, Englewood Cliffs, New Jersy, 1981.
3. P. B. Bochev and R. B. Lehoucq. On finite element discretization of the pure Neumann problem. Technical Report SAND2001-0733J, Sandia National Laboratories, P.O. Box 5800 Albuquerque, NM 87185, 2001.
4. Dietrich Braess. *Finite elements: Theory, fast solvers, and applications in solid mechanics*. Cambridge University Press, Cambridge, United Kingdom, 1997.
5. S. Brenner and R. Scott. *The mathematical theory of finite element methods*. Springer-Verlag, 1994.
6. F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*. Springer-Verlag, Berlin, 1991.
7. G. Carey and T. Oden. *Finite Elements. Computational Aspects*, volume 3. Prentice-Hall, Englewood Cliffs, New Jersy, 1984.
8. R. Cook, D. Malkus, and M. Plesha. *Concepts and Applications of Finite Element Analysis*. John Wiley and Sons, New York, NY, third edition, 1989.
9. Charbel Farhat and Michel Géradin. On the general solution by a direct method of a large-scale singular system of linear equations: application to the analysis of floating structures. *International Journal on Numerical Methods in Engineering*, 41:675–696, 1998.
10. V. Girault and P.-A. Raviart. *Finite element methods for Navier-Stokes equations*. Springer-Verlag, 1986.
11. P.M. Gresho and R.L. Sani. *Incompressible Flow and the Finite Element Method*. John Wiley and Sons, Chichester, England, 1998.
12. T. Hughes. *The Finite Element Method: Linear Static and Dynamic Analysis*. Dover, New York, 2000.
13. S.F. McCormick. *Multilevel Adaptive Methods for Partial Differential Equations*. SIAM, Philadelphia, PA, 1989.
14. J. Nocedal and S. Wright. *Numerical optimization*. Springer Verlag, New York, 1999.

15. M. Papadrakakis and Y. Fragakis. An integrated geometric-algebraic method for solving semi-definite problems in structural mechanics. *Computer Methods in Applied Mechanics and Engineering*, 190:6513–6532, 2001.
16. J. N. Reddy. *An Introduction to the Finite Element Method*. McGraw-Hill, New York, NY, second edition, 1993.
17. O. C. Zienkiewicz and R. L. Taylor. *The Finite Element Method: Fluid Dynamics*, volume 3. Butterworth-Heinemann, New York, NY, fifth edition, 2000.
18. O. C. Zienkiewicz and R. L. Taylor. *The Finite Element Method: Solid Mechanics*, volume 2. Butterworth-Heinemann, New York, NY, fifth edition, 2000.
19. O. C. Zienkiewicz and R. L. Taylor. *The Finite Element Method: The Basis*, volume 1. Butterworth-Heinemann, New York, NY, fifth edition, 2000.